RECORD OF SOCIETY OF ACTUARIES
1995 VOL. 21 NO. 3A

PHYSICIAN PERFORMANCE MEASUREMENT

Moderator: PATRICK J. DUNKS
Panelists: E. ANDREW BALAS*
RANDALL VOLLERTSON†
Recorder: TIMOTHY D. COURTNEY

Panelists will present ways to measure physician performance. These will involve analysis of historical data to monitor and evaluate performance of physicians and other providers.

MR. PATRICK J. DUNKS: We have a distinguished panel who will be speaking on what I think is a very hot topic. We're all trying to change the world in terms of how physicians practice medicine, and we're trying to get them to move toward efficient delivery systems. One thing that's necessary to improve performance is measuring performance. This issue is timely, new, and we'll see a lot more of it in the future.

Our first panelist, E. Andrew Balas, is an M.D. and has a Ph.D. in informatics. Dr. Balas is with Health Services Management in the School of Medicine at the University of Missouri, Columbia. He has presented numerous papers on measuring physician performance and authored many others.

Our second panelist, Dr. Randy Vollertson, is a senior physician consultant with Milliman & Robertson. Dr. Vollertson was a physician at Mayo Clinic and had worked on numerous projects related to measuring physician performance before joining Milliman & Robertson.

DR. E. ANDREW BALAS: Physician performance measurement is really one of the critical issues in health care. It's particularly becoming a critical issue given the emergence of managed health care. Before I begin, I would like to make you aware that there is a managed health care Internet discussion list. Feel free to subscribe; it's one of the few remaining things which is free on the North American continent. You can send an Internet message to subscribe to it. The Internet address is MHCARE-L@MIZZOU1./ MISSOURI.EDU

There's a good study that shows that there are a couple of things in patient care that should attract our attention and are somewhat unexplained. The study is by the Rand Corporation in California and it shows dramatic variation in utilization. Ten-thousand medical enrollees in 23 counties were included in the study. Three major procedures—coronary angiography, carotid endarterectomy, and upper GI endoscopy—were discussed. There was a huge variation in utilization. Now the next question, is this justified or not?

---

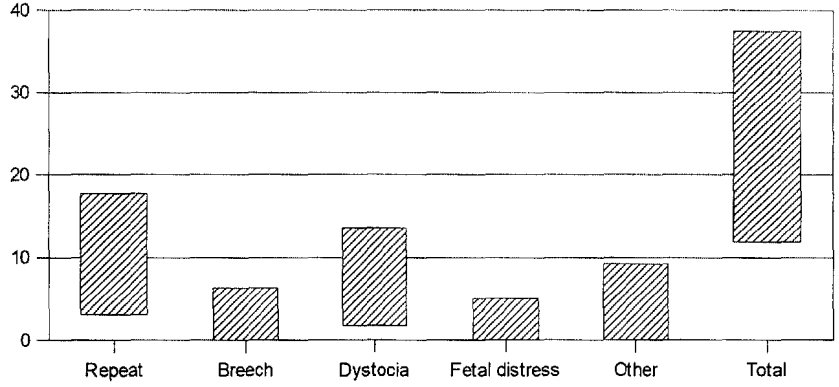
*Dr. Balas, not a member of the sponsoring organizations, is an M.D. and has a Ph.D. in informatics. He is at Health Services Management in the School of Medicine at the University of Missouri—Columbia in Columbia, MO.

†Dr. Vollertson, not a member of the sponsoring organizations, is a Senior Physician Consultant at Milliman & Robertson in Minneapolis, MN.

The same study evaluated the appropriateness of the procedures and again found huge variations. The most interesting and the most disappointing result of the study was that it cannot make the simple statement that inappropriate utilization explains high utilization rates and appropriate utilization rates explains low utilization rates. It's a very confusing issue. What we see is an ambiguity in panel evaluation of appropriateness.

Chart 1 shows the results of another study that looked at variations in C-section rates in the state of Missouri. With this study we can evaluate not only the total C-section rates, but we also get a little bit further by evaluating variation in selected categories. That says something more in terms of indications like repeat C-section rates or dystocia (inadequate progression of labor).

CHART 1
VARIATIONS IN C-SECTIONS
(PERCENTAGE OF ALL DELIVERIES)



This says that there's a disagreement among clinicians, and the variation is much larger. So if you look at the specific indications, the variation might be partially explained by the lack of agreement among clinicians.

A famous article in *The New England Journal of Medicine* stated: "Few medical issues are more controversial currently than carotid endarterectomy. But it's painfully obvious to experts on stroke that we still don't know which patients with what lesions detected by which test should be treated with what therapies." So there are a few uncertainties surrounding this therapy. This uncertainty comes not before introducing this procedure, but after paying for hundreds of such surgeries. You, as insurance people, know that's not what you would like to do, especially when the indications are so unclear. I have to tell you that after the publishing of this editorial, a clinical trial was started and that trial clarified the role of corotid endarterectomy. The indications are much clearer today then they were in those days.

How far are we from ideal care? That's an interesting issue. Fifty-five percent of primary care physicians reported that they perform the fundoscopic examination, an eye exam, on all their patients with insulin dependent diabetes mellitis (IDDM). Twenty-three percent report that they have performed a fundoscopic examination on less than half of their

patients with IDDM. You should know that almost every diabetic patient should get an annual eye exam. It's a critical procedure which can prevent loss of vision, a major complication of diabetes mellitus. It is recommended by the American Diabetes Association, and it is widely substantiated by various types of evidence. The problem is that only 55% of physicians perform this examination. So what is savings now will be a tremendous loss later. In addition, it indicates a deviation from the accepted or recommended standards of care.

Interestingly, a cross section in a national survey indicated that only 49% of adults who are diagnosed with diabetes reported having a dilated eye exam in the last year. Among people with retinopathy, only 61% had an eye examination. Frequently, there is a tremendous discrepancy between what physicians state they do, and what people report they receive. So, we have to carefully analyze whether or not the procedure was performed.

Next, I would like to discuss clinical practice variation. I would like to demonstrate the differences between various professions. A duck hunting example shows that a family practitioner sees a bird fly overhead and follows it with his rifle. Eventually, the duck flies off. The family practitioner says, "If you watch these things, they usually just take care of themselves." The internist sees the bird fly by and shouts, "There's a duck; rule out goose; rule out albatross, or swan; rule out whatever." By that time, the bird has disappeared, but you have to pay for all he's ruling out. The surgeon fires his gun and turns to the others and says, "Was it a duck?" Another surgeon fires at the bird and tells the pathologist go see if that was a duck. That's the careful person. The radiologist sees the duck fly up, jumps out of the boat and says, "I want another view of the duck. One more X-ray picture." The psychiatrist shakes his head and starts mumbling to himself, "Is it really a duck or does it just think it is, or do I just think it is?"

There are variations in clinical practice guidelines. Certainly you would like to get clear recommendations, and clinical practice guidelines are widely accepted sources of such recommendations. They are specifications for efficient care, developed by a formal process to incorporate published scientific evidence with an expert opinion. What is the hard core of these recommendations? Scientific evidence. What is the soft part? Expert opinion. What is the ratio of expert opinion based recommendations versus scientific evidence based recommendations in clinical practice guidelines? About 85% is expert opinion based and the rest is based on scientific evidence. So the hard foundation is not really there but what is there is something that might be useful. If it's 15% let's realize it and accept that it's 15%.

A guideline recommendation regarding estimation of post-void residual volume (PVR) states: "This test is recommended for all patients with urinary incontinence. Estimation can be made by abdominal palpation and percussion and/or on bimanual examination. When specific measurement of PVR is needed, it can be accomplished either catheterization or by pelvic ultrasound." This is from "Urinary Incontinence in Adults: Clinical Practice Guideline," Agency for Health Care Policy and Research, 1992. Estimation can be made by methods listed in the excerpt but this is a long recommendation for individual care. If you would like to measure physician performance, the question is, how can you measure based on this recommendation? There is nothing measurable here. Even the recommendation is somewhat unclear. So there's a recommendation, a clinical practice recommendation, and it's not really measurable.

There are other types of recommendations. For example, the overall C-section rate is supposed to be 15 or fewer per 100 deliveries. The primary C-section rate is supposed to be less than 12 per 100 deliveries. That's a much clearer, much more measurable recommendation. It's maybe not much better, but it's certainly much more measurable. Actually, it's interesting that the previous recommendation was developed by one arm of the U.S. federal government, The Agency for Health Care Policy and Research, and the other one was given by the Centers for Disease Control. So they follow different philosophies and, probably, the truth lies somewhere in between.

There are two major types of performance measures that are quality indicators in health care. The first is the outcome indicator, for example, case mix adjusted hospital mortality rate. The advantage is that it measures success in meeting patient needs. Patients would like to survive when they go to the hospital. Certainly that desire is measurable. On the other hand, the disadvantage is it doesn't say anything about the changes needed in the process. If you remember the story of the mortality statistics published by the Health Care Financing Administration, they say the mortality statistics were not continued because nobody could figure out what they should do if their mortality rate made them look bad. So that's certainly a serious disadvantage of outcome indicators.

The second type of performance measures are process indicators, for example, frequency and lengths of hospitalization. The advantage of process indicators is they quantitatively specify what needs to be changed—like the C-section rate. The disadvantage is they do not say anything about the effect and outcome. Why is 15% a good C-section rate? Why not 12%? Twelve percent has a much longer tradition.

Here's an example of a standard of care. A doctor prescribes a pure product, James E. Pepper Whiskey. It is the accepted standard of whiskey excellence from a medical standpoint, and most generally prescribed. But it's not the latest recommendation. It's from *The American Journal of Surgery* in 1907, or something like that.

What is the solution if you have the problem that there is a process indicator, and an outcome indicator? We would like to have something that can measure the outcome of care. Maybe we can use the overlap of the expert consensus and randomized controlled clinical trial (RCT) evidence. Randomized controlled clinical trials are really emerging as the ultimate tool to link process procedures to the outcome of care. If you have both the consensus of the RCT evidence and outcome, you are probably in good shape.

This is illustrated by the example of diabetic foot care. Diabetic foot care is something that is recommended regularly for every diabetic patient. The intervention target is 100% documented foot examinations and foot care education for diabetic patients. This is not rocket science type intervention, but it's very effective. The outcome is that patients receiving the intervention were less likely than controlled patients to have serious foot lesions and other dermatological abnormalities. The number one reason for amputation is complications of diabetes and, interestingly, simple foot care examinations and education can help to prevent a major portion of those amputations.

Another type of evidence is a guideline. For example, the feet should be examined at every regular visit for assessment of vascular status, state of skin condition and sensation. That is what the clinician would like to see—a recommendation for individual outpatient care.

Two critical types of evidence have been presented—direct evidence on outcome and expert consensus. Something else that might be helpful is miscellaneous evidence. Below is a hierarchy of evidence. It's a basic ladder that can help orientation in the vast amount of medical literature. There is a publication on everything and there are data published to justify just about anything in medicine. That certainly is a very disappointing situation, but if you have the right guides to identify the high-quality information and the high-quality evidence, then certainly you are in much better shape. You can rely on something that can be trusted.

Clinical Evidence Substantiating Practice Recommendations

A. Direct evidence on outcome
   1. Meta-analysis of randomized clinical trials
   2. Randomized controlled clinical trial
B. Expert consensus
   1. Clinical practice guideline
   2. Major review paper
C. Miscellaneous evidence
   1. Nonrandomized clinical study
   2. Epidemiologic study
   3. Cost-effectiveness study

The top-quality evidence, accepted by many, including the U.S. Preventive and Canadian Preventive Health Care task forces, is the direct evidence and outcome that can come from a randomized controlled clinic trial, or if there are several trials, then you can develop meta-analysis of randomized clinical trials.

The next category is expert consensus, which I have already discussed. Examples of miscellaneous evidence include clinical studies, epidemiological studies, studies showing the magnitude of the problem and cost effectiveness studies.

In a nonrandomized study at the University of Missouri–Columbia, several large clinical centers experienced a 44–85% reduction in the rate of amputation among individuals with diabetes after the implementation of an improved foot care program. Is this a miraculous result? The only problem is it's not true. The reason is that this study was not randomized. It was influenced most likely by serious types of bias and it is an overestimate of the real effect. However, there is an effect and randomized controlled studies have assured this. So if you would like to convince people and get some money for some programs, present these numbers. For example, an epidemiology study showed the magnitude of the problem—diabetic patients made up 39% of amputees and 42% of all operations. So, it's a frequent problem.

Next, I'd like to look at uncontrolled versus controlled studies. There's a famous example that looked at the electronic fetal heart rate monitoring. Early studies showed that the crude neonatal death rate was 1.7 times higher in unmonitored infants than those monitored. Thus, in the highest-risk group, 109 lives might be saved for every 1,000 babies born. Well, if you're not a murderer then you have to introduce this technology as soon as possible, because it's life saving. The only problem is this is not a randomized controlled clinical trial. It was also biased.

When you monitor this over time, there are various reasons why you can get better results when you measure things again, because, as science progresses, people pay attention and so on. In a multicentered randomized clinical trial, they concluded, compared with a structured program of periodic auscultation (that's the old fashion way), electronic fetal monitoring does not result in improved neurological development in children born prematurely. So, there is no effect. This triggered a very disappointing editorial from *The American Journal of Medicine* that said electronic fetal heartbeat monitoring failed to justify itself as a useful technology. So we spend a lot of money and we are actually getting nothing better than the traditional auscultation program.

I'd like to discuss interventions to change clinical practice. If you would like to measure performance, you have to have a reason to measure. The reason is to change and to make clinical practice more effective.

The physician's pen is the instrument responsible, in large part, for the high cost of health care. Physicians directly prescribe the provision of services that account for more then 70% of personal health care expenditures. We have to pay a great deal of attention to this pen if we would like to optimize care.

What are the ways of changing provider behavior in managed health care? You can present alternatives and show more cost-effective alternatives, you can use authorization (a very unpopular technique, but certainly it has some kind of effect), and financial incentives. Many believe financial incentives are the only tool. That's probably not true. Actually it's not even true that it has a beneficial effect in changing clinical practices.

What you see is that most of these techniques, such as concurrent reviews and education, rely on information. They present data on clinical practices. So information techniques seem to play a kind of central role. That's why we have to talk about the new role of information. In the early days, when Informatics was an emerging science in medicine, people believed that better information meant better decisions. What we experienced is that first, accurate information can be totally useless information. On the other hand, somewhat accurate information, or even inaccurate information, can trigger a beneficial change. So it's better to say, information is a clinical intervention which can and should change the practice of medicine.

Here's an example of an uncomplicated information intervention trial. The last mammogram date was to be displayed in the comments section of an encounter form. Of course, the purpose of this trial was to increase mammography screening. What was the effect? The percent brought up to date of those due at the start of the intervention period or who became due was 19.1% versus 12.1%. That was a 7% increase. Is 7% a miserably poor thing? No, that's real change. When you hear the comments from physician executives saying that when they want to make changes, they want 100% change, my reaction is of course, there is 100% change if you don't measure it. However, if you measure it, in reality it will be much smaller. That's what you see in randomized controlled clinical trials—the effect might be small but reliable. You should probably use that type of evidence rather than the great promises. Health care is really not the promised land; it's the land of promises. You have to be very careful.

If clinical trials are so important, we need to collect that type of evidence, if it's really top-quality evidence. That's what we are doing at the University of Missouri in Columbia. We collect randomized controlled clinical trials, using a design criteria of random or quasi-random allocation of intervention and contemporaneous monitoring of effect, informational utilization management intervention in the experimental group and no similar intervention in the control group and the effective measure of the process and/or outcome of the patient care. We want direct evidence on the quality of care rather than, for example, test results. Test results are like a multiple choice test and end up being a special continuing education program. They do not really have a direct relationship to the practice of medicine. We need direct evidence.

There's a wide variety of sources of clinical trials. Certainly they come from most countries. So there are world wide interests in this high-quality evidence.

What is nice is that publication of health services research trials is rapidly increasing, so we can get more and more high-quality evidence on interventions. There have been dramatic increases in all major categories—the computer-assisted information interventions, noncomputerized interventions, and utilization management type interventions. Theoretically, you can just go to the Medline database or to another major database and retrieve these clinical trials. That's the theory, but in practice, it's very difficult to find this type of evidence. There are several explanations. None of them makes us very happy. The reality is that we need a kind of systematic and widespread effort to find these clinical trials.

When you evaluate the quality of care or an intervention, you have to be very careful and comprehensive. Here's an example why. A consulting psychiatrist who is carrying out a patient satisfaction survey approaches a normally taciturn patient and asks him questions regarding his satisfaction with treatment. The patient states that the new medications are much better than the old medication that the psychiatrist prescribed. The psychiatrist was overjoyed at this. Before returning to his office he asks, in what way are they better? The patient replies that the old ones used to float when he flushed them down the toilet, whereas the new medicines disappear.

Let's look at an example of information intervention—physician reminder. Are physician reminders an effective information intervention and can they improve compliance for preventive health care measures? To answer this, our methodology was meta-analysis which ultimately combined data on over 5,000 patients in the tetanus immunization category and a similar number of patients in the cervical cancer screening category. We saw that the cervical cancer screening reminder effect was small. Something caused the effect to be small. The physician reminder was effective at improving preventive care practices, but the effect was limited.

Certainly clinical decision making is not the only factor that influences the practice of medicine. We have to focus on the entire process. The situation was somewhat different with the tetanus immunization trials, where it would clearly be unethical to conduct such a trial because there is enough evidence at this point in the literature to say that this is a highly effective information intervention. We can really recommend these. There is no reason to put somebody in the control group and withdraw this highly effective reminder intervention. Simple information can change the practice of medicine.

The intervention of peer comparison feedback is one of the favorites of the utilization management people. Can peer comparison feedback, physician profiling and effective information intervention change the average frequency of targeted clinical activities? To look at this, we used meta-analysis which involved 550 physicians. What we saw was a small effect. Physician profiling was statistically significant, but had a small effect on utilization. Actually the effect was so small that it's questionable that it can justify the process of creating these feedback reports. Which utilization management is most common in the U.S.? Fifty percent of U.S. physicians are subjected to physician profiling and the effect is very small. There is some analytical evidence that if you provide the comparative data then the low utilizers will go up, the high utilizers will go down and your average will stay the same, which is not your purpose.

We have to reevaluate physician profiling as a technique. What can we do better? Maybe we should improve our reports. A way to do that is to look at the comparative data. A clinical trial funded by the Missouri Kidney Program presented national, Missouri, and dialysis-center-specific data in a simple report, on the number of patients and percentage of patients allocated to peritoneal dialysis. You should know that when the kidney fails there is no other solution other than dialysis or kidney transplant. Dialysis is really an unavoidable first step for many people. The question is what type of treatment should be selected. There are two major competing treatments. The first one is hemodialysis, which is more expensive. The second is peritoneal dialysis, which is less expensive. Historically, peritoneal dialysis was considered an inferior or somewhat second class treatment, because the results were not as good. Today, this is no longer true. The results are equal and in some instances actually better. We need to encourage this type of intervention; That's the purpose of peer comparison feedback. But that's not enough. We can add scientific evidence through a direct mail concept report. We simply say to the physician, "That's what you do; we don't make any statement; we don't judge you." Making a judgment would be very unpopular. We don't take that risk. In addition to presenting data, we provide the scientific evidence pertinent to the selection of this therapy. In this case, it would say you should consider peritoneal dialysis more frequently and you should try to allocate more patients to this treatment modality.

What we found in this study was that the number of patients assigned to peritoneal dialysis was 14.3% versus 2.4% in the control group. That was a statistically significant result. The accrued savings for the first month of treatment was close to $8,000. That's certainly something for us to look at.

When you measure efforts to influence people an issue is that you're interfering with medicine in the interest of business success. However, when we evaluate information intervention, we have to realize that they move in the same direction. We should focus on these information interventions. The best example is preventive care. Increasing preventive care can prevent costs down the road, and it's also good medicine. As an example, I'll use an analysis related to the use of antibiotics from a famous experiment at Intermountain Health Care in the Utah and Nevada area. In this experiment, they sent reminders to physicians that patients require antibiotic prophylaxis during surgery because, in some cases, it's just omitted—that's the nature of the process and human interaction. The result was fewer wound infections and also, reduced cost of care to prevent infections. That is certainly something that can be very beneficial.

Another example is recalling patients for cancer prevention. On one hand, you get cancer prevention, which is a clinical activity. On the other hand, it is service and marketing. And it's good business and good medicine at the same time.

Another famous experiment at the University of Indiana, which produced a very large number of randomized controlled clinical trials, looked at displaying charges for a test. They display charges for tests, and that simple intervention, without making any firm recommendation or being formal or disappointing people, reduced the number of tests. Fewer tests were ordered, so it's a good cost-saving opportunity, and the clinician made an informed decision—certainly something that constitutes good medicine. There are ways to combine these things. Certainly, that would be the direction that we should go.

DR. RANDALL VOLLERTSON: I'm a physician with the Minneapolis office of Milliman & Robertson. To compliment Dr. Balas' talk, I'm here to talk about how you might look at some of these physician performance measures and apply them to your own situation.

First of all, I'd like to give you some practical guidelines to interpret these performance measures, since this can be tricky. Second, I'd like to talk about the currently publicly available performance measures. Many of these are considered proprietary and hard to get. As you will see, a lot of the information turns out to be the same. Third, I'm going to give a couple of examples. The first example is from a large individual practice association (IPA) health maintenance organization, the second example is the health plan employer data and information set (HEDIS) project. HEDIS is actually a large project looking at clinical quality indicators. There are three subprojects that evaluated health plans: one was in New England, one was in California, and there was a national project that involved 21 plans. They're all similar, so I'm only going to talk about the national project. Finally, I want to give you a couple of snapshots of four studies that point out some of the problems that may be inherent in evaluating physician performance.

When one begins to look at and interpret physician report cards, there are several things to keep in mind. I think the first thing we want to keep in mind is, what's the purpose of the card? Is it to select or retain providers? To monitor them? To change their behavior? Is it to report externally or is this for marketing purposes? If you're just giving providers some feedback, the rigor that may be required is a lot different than if you want to select or retain them. Sometimes providers will hire attorneys to point that out to you so you need to be aware of that. I'm told they've been known to do that.

The second thing to keep in mind is, who is doing the evaluating? Is the evaluating entity calling a national press conference to announce their immunization rate? That may not have as much credibility to some people as a report by a government agency, a payor, or an independent organization. I think that you can appreciate that each of those evaluating groups has different implications.

The next thing to keep in mind is, what entity is being evaluated? Are we looking at a health plan, a hospital, a physician group, or an individual physician? As you proceed down that list the numbers become smaller, and you get into some small number problems. Also, the information becomes harder to get. I'm sure one would think nothing of seeing public insurance companies' reports of how they did each year. In fact, that's required if they're traded on the stock exchange. Regional reports may be OK. I doubt if you'd want

to read individual performance reports in the paper; maybe you would, I don't know. You'd probably want to read everybody else's but not your own. Health plans are the same, and as they move from different hospitals to individual physicians the information is kept a little more confidential.

What are the sources of data? There are really two broad categories: administrative and clinical. They each have different implications for accuracy and cost. As you proceed from administrative to clinical data, the accuracy tends to decrease and the cost tends to increase. You might keep in mind that the denominator of your report cards can vary. If you're comparing data it's important to compare data with the same denominator.

What are the types of report cards? Are they just a noncomparative statement? Do they report on just the number of mammograms last year, or do they compare in a crosssectional way against either a norm or peers. The HEDIS report is one of the latter. Are they a longitudinal comparison? Do they compare what happened this year to last year? Are we improving? Are we getting better? Are we getting worse?

Finally, one can categorize the reports from these report cards in several ways. Here's one I like to use. You can look at financial parameters, which may be either monetary or nonmonetary proxies. You can look at quality measures, which can include structure, process, outcome, and they can be both administrative and clinical. You can also look at utilization, which may be thought of as a nonmonetary process measure; it is one of the most popular, particularly on the cost side.

Some examples of measures include financial measures of the cost per case with the claims per year. Structural quality might be defined as board certification of your panel—whether your plan or the hospitals have a certification or how many hours a practice is open. Process measures are frequently chosen because they're common, because oftentimes primary care physicians are the ones that are measured, and because they are relatively easy to come by compared to other clinical data. So screening tests, immunization data and operating room times are all key measures.

Administrative outcomes are also important. You can measure examples or patient satisfaction and disenrollment. Dr. Balas' story about the satisfied patient who flushed his medicine down the toilet reminds us all that these aren't exactly the final outcomes.

Clinical outcomes, which are a little more difficult to measure, are really what we're interested in from the standpoint of the care. Mortality is common. I call that the big outcome. It's one that's easy to measure. Morbidity and others are much less common and much more difficult to measure. Readmission rates, infection rates, and intensive care unit (ICU) transfer rates are generic examples of these. We're all familiar with these utilization measures; these are quoted by almost everybody. The health care industry today brags about its length of stay.

To summarize, the requirements of a report card are to have reliable data, to have a valid measure of that data, and to have a valid analysis of it. Valid analysis is something that's statistically and clinically significant. I think we saw examples in the preceding discussions of things that were perhaps statistically but not clinically significant. You'd also like things

to be actuarially adjusted. Ideally you'd like them to be adjusted for the severity of the clinical case. I'm going to give you some examples of that a bit later.

Next, I'd like to survey the current market—in other words, structure, utilization, and administrative outcomes, primarily. Much of the clinical outcomes focus on preventive care. We're beginning to see a great deal of substantial other clinical outcomes. However, when you look at these in a little more detail, and you look for nonself-reported clinical outcomes, there are really only four reports that are currently widely and publicly available. These are all hospital based. They focus on mortality, utilization and other procedures. So I think we're still in the infancy of public reporting of the physician and other provider performance.

Now I'd like to turn to two examples. The first is going to be that of a large HMO. I'm going to focus on the report cards for the primary care providers. I'll tell you that the specialists have a little different reporting system. It's not nearly as involved in this organization. I'm not going to touch on that. I'll also tell you that the hospitals have a rather similar reporting or evaluation system.

This organization wants to look at both quality and utilization. They want to link this to the financial compensation. Now, this information doesn't come cheap. They designate a quality person in each office, who spends part of his or her time evaluating this. They have a medical director that they say visits annually. I've heard from others that the medical director visits more frequently than just annually in some cases. They have an elected advisory quality assurance committee as well.

Their review is divided into three parts, which they call quality review, comprehensive care and utilization. The first two are not straightforward. Let me explain each of those in a little more detail. The quality review consists of four parts. There's a survey of the patients. There is a focused medical care review, wherein, they look at the medical records and compare them to nationally recognized (whatever those are) standards. I can tell you nationally recognized standards are all over the place. They also look at the transfer rate. Finally, they look at what they call managed care philosophies, which is code talk for getting along with the HMO.

What they call comprehensive care consists of several things. It consists of access. They look at three parameters there, office hours, the number of services offered, (remember these are primary care physicians, so they want to encourage them to offer more services), and their acceptance of catastrophic cases (those are time consuming, costly cases). We'll talk about catastrophic cases a little more because they do make a correction for that as I'll explain.

They also look at the completion of continuing education—whether your practice grows and whether you're linked up to headquarters on the computer. For utilization there are pretty standard measures. Hospital care is measured in days per thousand per year. Specialty and emergency room costs are measured in per member/per month dollars. You are compared within your own specialty.

Now, they link financial rewards in the following way. Everybody gets a base capitation that requires a minimum quality and utilization standard that's reassessed every six months.

In addition to that, you can get some extra financial awards if you are one of the more successful members on your evaluation. They look at hospital, specialist and emergency room utilization. You can get additional distributions and increases in your capitation if you're at the top of the class. These are adjusted for quality and they are also adjusted for comprehensive care. Finally, you also get some status payments if you still have an open practice.

In adjusting this for the burden of illness, they actually have a fairly sophisticated technique. They reason that age and sex may determine the prevalence of a disease, and that utilization is, in turn, a function of that. They also realize that physician selection is not random, so they try to make an additional adjustment for case mix. They chose an empirical approach to this. They felt that total cost was a function of the referrals, referrals being both hospital and consultative. They also felt that the efficiency was a function of the rate in referral and cost of referral. So, they selected some economically important conditions and aggregated these into episodes by reviewing not only claims reports, but also encounter data. They also looked at each specific site. From this they derived a principal diagnosis for each patient and aggregated these for each office. So, an office could be a solo practitioner, or a group of five or ten. Then they looked at the number of these principle diagnoses compared to the expected number that were capitated for a given diagnosis to get an idea of what the burden of illness for each office was. This then went into their formula for determining the extra awards. What they found is, at least among their primary care physicians, the rate of referral was predictable, but the cost of referral was not.

The second area I'd like to talk about is the HEDIS national survey. It illustrates some of the clinical parameters and utilization parameters that are being used today. It consisted of 21 health care delivery organizations and was a voluntary survey. They covered the entire U.S. It was supervised by the National Committee for Quality Assurance, and it was a subset of HEDIS 2.0. They looked at five areas: member satisfaction, quality and access, the physician network, utilization, and membership and finance. I'll give you some examples.

First, there is quite a range among the 21 plans for member satisfaction. Second, for quality and access, we get into some more clinical process measures. There was quite a variance in immunization rates and mammographic screening. Diabetes eye examination rates range from 12% to 58%. All but the best plans were still under half. Asthma admissions rates and low birth weights were summarized also. You might ask, are some of these fair? Maybe. Maybe not. I think these need to be actuarially adjusted, and there are probably case mix factors here.

They looked at the physician network and found in some of these plans that 17% of the doctors quit every year. That's not good. They look to some structural measures there. For utilization, there is quite a range. There is a wide variation in coronary bypass utilization.

Finally, they looked at what they called membership and financial measures. They looked at the disenrollment rates which were sort of high for many established plans. They were name brands; you'd recognize them. In fact, you can buy the report. They have statistical

comparisons of individual plans in the report. Finally, they looked at some of the financial parameters such as the loss ratio and premium rates.

Let me just give you a couple of caveats. I'm going to go over four studies, very quick snapshots. I'm not going to get into the methods. The first one is regarding case-mix adjustments. They looked at The Harvard Community Health Plan—a staff model HMO. They looked at 52,000 patients and 52 practices. They derived what they called, standardized referral rates. These are referrals for consultations. They adjusted this for age and sex and then further adjusted it for ambulatory care groups, which incorporate ambulatory diagnostic groups, which is category of chronic disease. Then they further adjust the ambulatory diagnostic groups for age and sex. By doing this, they get ambulatory care groups.

When they ran a multiple regression analysis on the predictors of the referral rate (that's the unadjusted referral rate), they identified the physician's age, the years in the plan, the years in practice, the full-time employees and a couple of sites as the parameters. They found an association between case mix and practice intensity. In other words, the people that saw more patients per hour had more referrals. That was the only significant predictor. So you might want to think about allowing for that.

A large study looking at over 15,000 physicians in two states looked at the resource-based relative value schedule (RBRVS) per hospital stay. So they're looking at inpatient parameters here. They found out the Florida physicians had lots more than Oregon physicians. In fact, this is true. They were able to take this data, and were able to zero in on different hospitals. They found differences within each state. They were also able to further zero in on individual providers and identify the specific physician outliers. They were also able to look at different procedures. In fact, procedures at Site A may use more endoscopy (that's how they ran up their RBRVS values), while physicians at Site B used less endoscopy but more imaging. So not only do things get to be physician and site specific, but they also get to be disease and procedure specific. So you might not want to fire your best physician at Procedure A, just because he or she is bad at Procedure B; maybe you just want to redirect your efforts or focus them. So report cards can get to be tricky.

Finally there's the big question of value. They looked at 12 diagnosis related group (DRGs) compared to faculty and community staff in a large university hospital. They looked at experience for one year. They found that the faculty was clearly more expensive than the community staff. They also found that faculty staff had fewer deaths per 100 admissions than the community staff. This death difference disappeared after one year. Is that worth it? I don't know. I don't think that's an actuarial question. That's something for philosophers and priests and perhaps economists to debate. At any rate, I hope I've gone over some of the practical aspects of report cards.

In conclusion I'll say, I think report cards do require careful interpretation. I think we're going to see much more of them. There are more starting to appear. There's an increasing focus on clinical outcomes versus the cost, or what I call value—what you get for your dollar. I also think the reports are showing increased quality.

FROM THE FLOOR: My question has to do with why it's taking so long to compile data. We've been in the computer age now for at least a couple of decades. It seems like the medical industry is so far behind. I'm just wondering why academic institutions haven't at least tried to collect data, so we'd at least have databases, regardless of whether or not we'd use them for physician profiling.

DR. BALAS: I think it's a very good question. Why is evaluating practice patterns still the crisis de jour for most executives and not a routine process or a natural output of the information systems. Most of the health care information systems are not prepared to deal with groups of patients. They are increasingly better at supporting individual patient care. At the same time, they are unable to provide the fundamental data. The HEDIS indicator is a very good example. Diabetic annual eye exam. There are several problems. How many diabetic patients are in the plan? How many of those patients had an eye exam? How many of those did not have an exam during the previous year? So these are very simple questions, and the answer is usually in chart review. It's a shame that information systems currently are inadequate, and I think that they increasingly realize that there is much more attention needed for the group analysis of practice patterns. I hope that there will be a change, but most of the current information systems are just unable to produce the results.

DR. VOLLERTSON: I'd agree. I would add that these data are very expensive to get these days. This HEDIS study was. I don't know how much it cost, but I think you're talking over six and probably seven figures. And that's one year's data from 21 plans. The other thing is, I don't think the tools have been there in the past. I suspect we'll see much more data gathered with distributed network computing as people get into organized plans. Many people in the commercial world don't want to underwrite this gathering of data even when, in the long run, it may save money. It's not in anybody's budget.

MS. JOAN P. OGDEN: So many of the reports that I see in the various journals crossing my desk indicate that telling a physician something is not very effective in changing practice patterns. We've been talking about telling physicians. Do we need to use sticks?

DR. BALAS: We are biased, but we say no. Actually, I disagree somewhat with the statement that telling physicians would not help. Actually, it is an effective intervention. We should just appreciate the 5–10% improvement in quality. That is an achievement. It's very much like when you invest. In what would you like to invest? Something which produces a 5% or 10% return, or something that could produce a 100% return, but maybe 100% loss? I think that the 5–10% change should be acceptable. We have effective information for telling physicians the interventions, which make a difference.

DR. VOLLERTSON: I have two responses. The first is with regard to telling physicians. I'll say it doesn't work by itself, but it does work better then not telling them. In fact, education is helpful, but you need a community leader or a respected physician to lead the effort. That helps a lot. I think you did need to reinforce it with continued feedback. That seems to help some. And you need to continue that feedback. Most industries find that positive rewards work better then negative ones. As I said, there are a number of problems here. Most of these data aren't actuarially adjusted, and it isn't adjusted for case mix. You can manage all the site A/site B, older doctors/younger doctors stuff when, in fact, that might not be the problem if you don't have the right data. So until you have good data, I'd say you might do more harm than good. I would say that only telling physicians is not

good, but it's the first step as part of an educational process. I think you do need to put positive and negative rewards in place, and I favor the positive ones myself. I can tell you that as one who's done it, dealing with physicians can be very difficult. They can act out or worst of all, they can go limp, and then you're really stuck.

MR. THOMAS L. HANDLEY: I noticed, as both of you made your presentations, you were looking at and measuring many different values. It didn't appear that we received much concrete feedback or things that, as an actuary, we're used to looking at (statistics). We couldn't put a value to them. Many of the things that were being measured or we're attempting to measure look like we're just barely going down the road of performance measurement. The things we're looking at are getting more complicated and more detailed. We can spend, as an industry or as a society, millions of dollars on trying to measure these things and find out that we're not getting anywhere. It's like trying to nail Jello to a tree. You can try and hold on to it, but it just keeps slipping around the nail. Are we making things more complicated? Should we just focus on the simpler things that are very definable and measurable? Should we just focus on what our objective is, and that's trying to keep the cost of health care at a reasonable level.

DR. BALAS: Good point. Quality measures are quite confusing. There's a wide selection of quality measures. I thought the same thing when I purchased my car. I read the complete car cost guide, and I was completely confused at the end. Certain things were good, some other things were bad, and who knows what is more important. In health care it can be very much the same. There are two types of quality indicators. One is the overall types of quality indicators, and those are recommended to be severity adjusted. The other type is quality indicators which are like tracers. They select a very specific aspect of care, like a diabetic eye exam. Why is a diabetic eye exam more important than a flu vaccination? I don't think that anybody can answer that, but some of these things indicate overall quality. We probably need an appropriate mixture of these two types of variables, and I agree that measuring quality is an evolving science. To spend money without knowing what you get, I think, is very dangerous. I don't feel that is an alternative.

DR. VOLLERTSON: I think you've raised some very good points. I'll point out a couple of things. First of all, this is what's out there now. Second, it's a lot better than what used to be out there. I think we are making progress. Third, I think we'll probably never be able to pin everything down. We won't ever be able to tell whether a 58-year-old man who has a prostrate nodule, has diabetes mellitus, had a heart attack six months ago, and has a family history of rheumatoid arthritis should or should not have a prostate biopsy. That gets to be too narrowly focused. The numbers just aren't there.

There was a very good session about risk adjustment. Essentially, the best actuarial predictions for the year-to-year risk adjusters are R^2's of 0.09–0.11. Some things are too uncertain to be measured in every case. However, I still think they are worth doing, because an R^2 of 0.09 or 0.11 may not be much in physics or electronics, but I'd sure like to have it in the stock market.

MR. STEELE R. STEWART: HMOs have been progressively bringing down utilization like hospital days and length of stay through some management techniques. The focus of your talk has been more on the physician side. With regard to looking ahead to the long

term, what kind of change do you see as far as the utilization rates, on the physician side, or the cost of care on the physician side over the next five to ten years due to some of these things that you're bringing in?

DR. VOLLERTSON: What does the future hold? Well, I don't know. But I'll say this, I think you're going to see more efficient practice. We're going to have better measures of value. I think physicians will get more interested in this. That's what you want, as far as efficient care. It might not be what those of you who work for insurers want, because I think you'll find them to be your partners instead of your contractees. I think the systems will get better.

DR. BALAS: I think HMOs tout the principle that the empty bed is the happy bed. And certainly one of the major shifts I would expect is the continuous downsizing of the inpatient facilities and a very strong shift towards outpatient care and coordinated outpatient care. The problem is that everybody does a perfect job, but care is still totally uncoordinated and full of waste. I think that coordination will be more of an issue than we perceive it to be today.

MR. WESLEY S. CARVER: I think probably the majority of managed care plans have some sort of financial incentives for the physicians. It may be a withhold or referral and hospital services fund. I was wondering, as physicians, what your thoughts would be on the effectiveness of financial incentives, both with and without some quality experience data feedback.

DR. VOLLERTSON: Well, I personally worked at the Mayo Clinic, where everybody gets a salary—no withholds, no bonuses, just salary. I can tell you that the utilization in the division of 21 doctors that I managed was 60% more for the highest person compared to the lowest. I don't mean utilization, I mean the amount of work they did. I can tell you that within the same departments people getting the same salaries did totally different numbers of tests in some cases. So this isn't all financially related. Now, I also think any physician who tells you that they don't respond to money is not telling you the truth, because almost all of them do. I've said before I think the same *X*%, (and I hope it's a small percent of physicians who will do extra tests in a fee-for-service system) that will tell you that your upset stomach might be angina and order a thallium scan, are probably the same 15% who, in a capitated system, would tell you your early angina might be just an upset stomach. Keep in mind that those dollars show up every month, whereas the consequences of somebody running around with angina may never show up at all. That's the challenge I think that all of us—payors, providers, and intermediaries—have to face. So I do think financial parameters work. I think they have to be used with care. I think if they're not coupled with quality, you're asking for trouble. The trouble is you may not know about it until it's too late. I think doctors respond to things other than financial parameters.

DR. BALAS: I think your question implied that physicians would like to get quality data. I would very much agree with that. People would like to get information. Certainly, they don't want to get it in a punishing way. But to inform is very important, and sooner or later we will need to provide more accurate information to the physicians to improve their practices. Nobody wants to be an outlier. There's a tremendous desire to become an average among most people.

MR. RICHARD E. ULLMAN: So far all the data that has been presented I assume is being used by carriers or employers, or large purchasers of care. Do you think there will ever be a day when the data will be available to individual patients? Will these data show that large purchasers of care will tend to select the most efficient physician from a cost standpoint, or will the patient want the guy or gal who does more tests, because he or she feels safer with that physician?

DR. BALAS: Actually patients should be users of quality data. That's certainly a challenge. That's one of the problems of case-mix adjustment, because that's not what you buy when you go to see your physician. A patient wouldn't say, I would like to get a case-mix-adjusted hospitalization rate. That's not how you express yourself. You would like to get good quality care and the quality should be the same for you as for your physician. That's why the tracers like the diabetic eye exam can be very good. If you are informed that a diabetic eye exam is important for you, your physician should get the same information. You are working towards the same goal. That's what is called shared patient/physician decision making, which is an increasingly popular concept. Certainly that should encompass the quality measurement as well.

DR. VOLLERTSON: That is a good question. Actually patients do, in many cases, seek physicians who do what they want. Unfortunately, they're not always informed. For example, I would venture that a lot of people would really rather not have their eyes dilated and have that exam. Ophthalmologists decide. Many physicians don't think about that. Either they're not informed or they have no immediate thought of it. So, to rely solely on the patients to get their care may not be the best solution. They may think the drugs work fine because they flush nicely. That may be a satisfied patient. But it's probably not good quality of care.